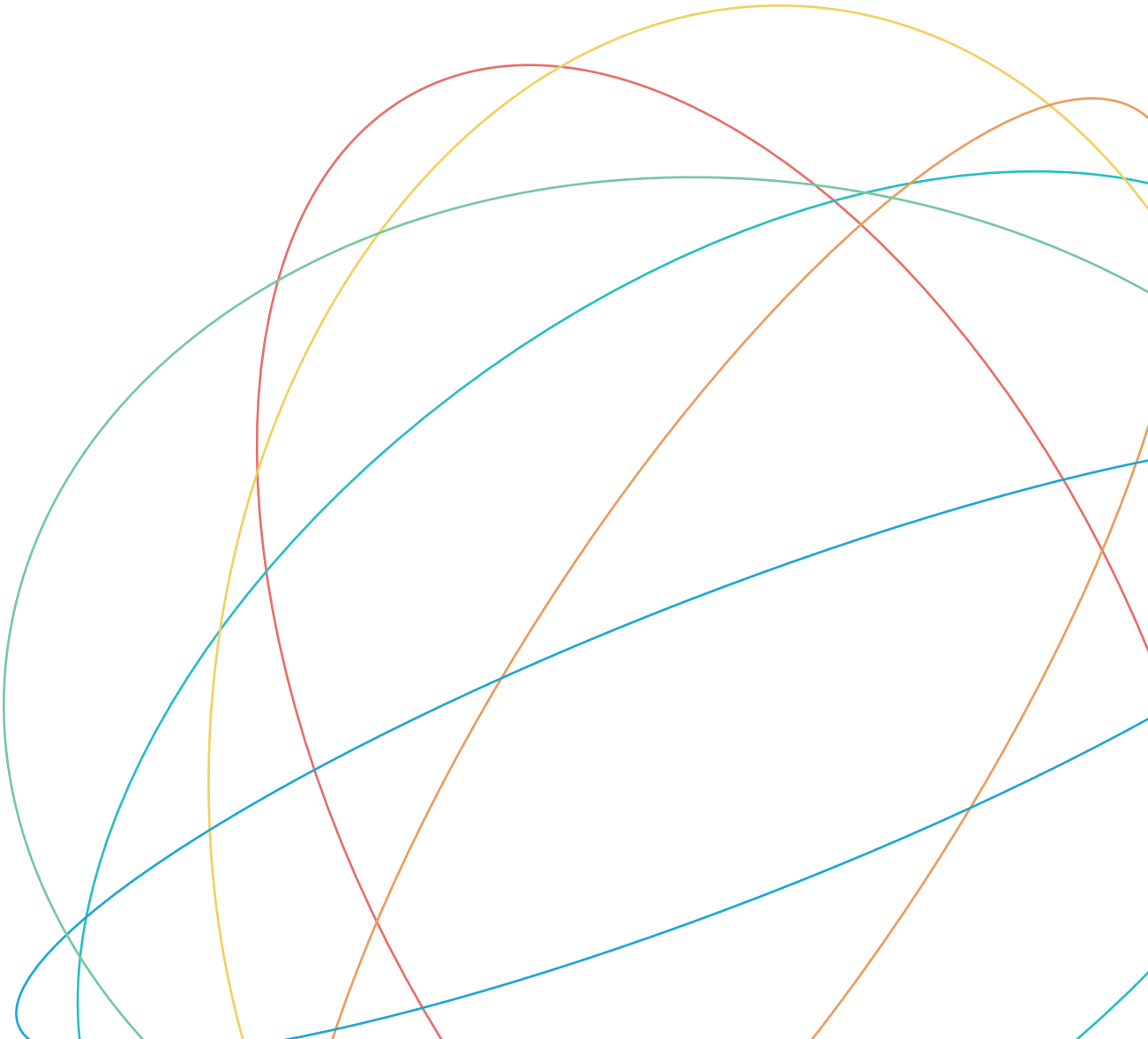
HUAWEI

AI安全白皮书



执行摘要

近年来，随着海量数据的积累、计算能力的发展、机器学习方法与系统的持续创新与演进，诸如图像识别、语音识别、自然语言翻译等人工智能技术得到普遍部署和广泛应用，人工智能正朝着历史性时刻迈进。与此同时，AI对于传统计算机安全领域的研究也产生了重大影响，除了利用AI来构建各种恶意检测、攻击识别系统外，黑客也可能利用AI达到更精准的攻击。除此之外，在关键的AI应用场景上，AI自身的安全性变得前所未有的重要，极需要构建一个不会被外界干扰而影响判断的健壮AI系统。可以说**AI帮助了安全，安全也能帮助AI**。

本白皮书主要目的是探讨AI自身的安全，确保AI模型和数据的完整性与保密性，使其在不同的业务场景下，不会轻易地被攻击者影响而改变判断结果或泄露数据。不同于传统的系统安全漏洞，机器学习系统存在安全漏洞的根因是其工作原理极为复杂，缺乏可解释性。各种AI系统安全问题（恶意机器学习）随之产生，闪避攻击、药饵攻击以及各种后门漏洞攻击层出不穷。这些攻击不但精准，而且对不同的机器学习模型有很强的可传递性，使得基于深度神经网络（DNN）的一系列AI应用面临较大的安全威胁。例如，攻击者在训练阶段掺入恶意数据，影响AI模型推理能力；同样也可以在判断阶段对要判断的样本加入少量噪音，刻意改变判断结果；攻击者还可能在模型中植入后门并实施高级攻击；也能通过多次查询窃取模型和数据信息。

华为致力于AI安全的研究，旨在提供一个令用户放心的AI应用安全环境，为华为AI使能构建智能世界的新时代愿景与使命做出贡献。为了应对AI安全的新挑战，本白皮书提出了将AI系统部署到业务场景中所需要的三个层次的防御手段：**攻防安全**，对已知攻击设计有针对性的防御机制；**模型安全**，通过模型验证等手段提升模型健壮性；**架构安全**，在部署AI的业务中设计不同的安全机制保证业务安全。

未来，华为的AI安全任重而道远。在技术上，需要持续研究AI可解释性，增强对机器学习工作机理的理解，并构建机制性防御措施搭建AI安全平台；在业务上，需要详细剖析AI在产品线的应用案例，落地经过测试和验证的AI安全关键技术。以“万物感知、万物互联、万物智能”为特征的智能社会即将到来，华为愿与全球的客户和伙伴们共同努力携手并进，共同面对AI安全挑战。

目录

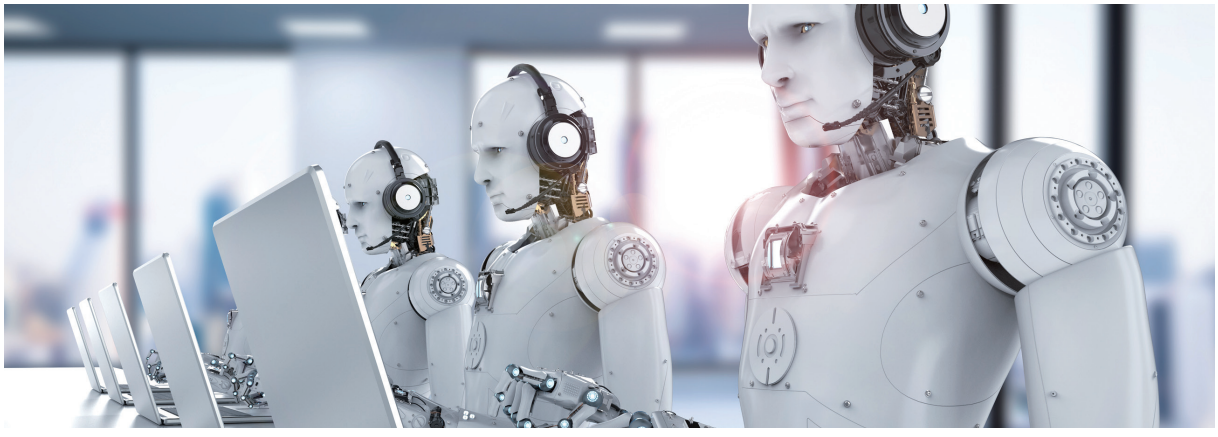
01 迈向智能社会

近年来，随着海量数据的积累、计算能力的发展、机器学习方法与系统的持续创新与演进，诸如图像识别、语音识别、自然语言翻译等人工智能技术得到普遍部署和广泛应用。越来越多公司都将增大在AI的投入，将其作为业务发展的重心。华为全球产业愿景预测：到2025年，全球将实现1000亿联接，覆盖77%的人口；85%的企业应用将部署到云上；智能家庭机器人将进入12%的家庭，形成千亿美元的市场。

人工智能技术的发展和广泛的商业应用充分预示着一个万物智能的社会正在快速到来。1956年，麦卡锡、明斯基、香农等人提出“人工智能”概念。60年后的今天，伴随着谷歌DeepMind开发的围棋程序AlphaGo战胜人类围棋冠军，人工智能技术开始全面爆发。如今，芯片和传感器的发展使“+智能”成为大势所趋：交通+智能，最懂你的路；医疗+智能，最懂你的痛；制造+智能，最懂你所需。加州大学伯克利分校的学者们认为人工智能在过去二十年快速崛起主要归结于如下三点原因[1]：1）海量数据：随着互联网的兴起，数据以语音、视频和文字等形式快速增长；海量数据为机器学习算法提供了充足的营养，促使人工智能技术快速发展。2）高扩展计算机和软件系统：近年来深度学习成功主要归功于新一波的CPU集群、GPU和TPU等专用硬件和相关的软件平台。3）已有资源的可获得性：大量的开源软件协助处理数据和支持AI相关工作，节省了大量的开发时间和费用；同时许多云服务为开发者提供了随时可获取的计算和存储资源。

在机器人、虚拟助手、自动驾驶、智能交通、智能制造、智慧城市等各个行业，人工智能正朝着历史性时刻迈进。谷歌、微软、亚马逊等大公司纷纷将AI作为引领未来的核心发展战略。2017年谷歌DeepMind升级版的AlphaGo Zero横空出世；它不再需要人类棋谱数据，而是进行自我博弈，经过短短3天的自我训练就强势打败了AlphaGo。AlphaGo Zero能够发现新知识并发展出打破常规的新策略，让我们看到了利用人工智能技术改变人类命运的巨大潜能。

我们现在看到的只是一个开始；未来，将会是一个全联接、超智能的世界。人工智能将为人们带来极致的体验，将积极影响人们的工作和生活，带来经济的繁荣与发展。



02 AI安全面临五大挑战

AI有巨大的潜能改变人类命运，但同样存在巨大的安全风险。这种安全风险存在的根本原因是AI算法设计之初普遍未考虑相关的安全威胁，使得AI算法的判断结果容易被恶意攻击者影响，导致AI系统判断失准。在工业、医疗、交通、监控等关键领域，安全危害尤为巨大；如果AI系统被恶意攻击，轻则造成财产损失，重则威胁人身安全。

AI安全风险不仅仅存在于理论分析，并且真实的存在于现今各种AI应用中。例如攻击者通过修改恶意文件绕开恶意文件检测或恶意流量检测等基于AI的检测工具；加入简单的噪音，致使家中的语音控制系统成功调用恶意应用；刻意修改终端回传的数据或刻意与聊天机器人进行某些恶意对话，导致后端AI系统预测错误；在交通指示牌或其他车辆上贴上或涂上一些小标记，致使自动驾驶车辆的判断错误。

应对上述AI安全风险，AI系统在设计上面临五大安全挑战：

- **软硬件的安全：**在软件及硬件层面，包括应用、模型、平台和芯片，编码都可能存在漏洞或后门；攻击者能够利用这些漏洞或后门实施高级攻击。在AI模型层面上，攻击者同样可能在模型中植入后门并实施高级攻击；由于AI模型的不可解释性，在模型中植入的恶意后门难以被检测。
- **数据完整性：**在数据层面，攻击者能够在训练阶段掺入恶意数据，影响AI模型推理能力；攻击者同样可以在判断阶段对要判断的样本加入少量噪音，刻意改变判断结果。
- **模型保密性：**在模型参数层面，服务提供者往往只希望提供模型查询服务，而不希望曝露自己训练的模型；但通过多次查询，攻击者能够构建出一个相似的模型，进而获得模型的相关信息。
- **模型鲁棒性：**训练模型时的样本往往覆盖性不足，使得模型鲁棒性不强；模型面对恶意样本时，无法给出正确的判断结果。
- **数据隐私：**在用户提供训练数据的场景下，攻击者能够通过反复查询训练好的模型获得用户的隐私信息。

03 AI安全典型攻击方式

3.1 闪避攻击

闪避攻击是指通过修改输入，让AI模型无法对其正确识别。闪避攻击是学术界研究最多的一类攻击，下面是学术界提出的最具代表性的三种闪避攻击：

对抗样本的提出：研究表明深度学习系统容易受到精心设计的输入样本的影响。这些输入样本就是学术界定义的对抗样例或样本，即Adversarial Examples。它们通常是在正常样本上加入人眼难以察觉的微小扰动，可以很容易地愚弄正常的深度学习模型。

微小扰动是对抗样本的基本前提，在原始样本处加入人类不易察觉的微小扰动会导致深度学习模型的性能下降。Szegedy等人[2]在2013年最早提出了对抗样本的概念。在其之后，学者相继提出了其他产生对抗样本的方法，其中Carlini等人提出的CW攻击可以在扰动很小的条件下达到100%的攻击成功率，并且能成功绕过大部分对抗样本的防御机制。

物理世界的攻击：除了对数字的图片文件加扰，Eykholt等人[3]对路标实体做涂改，使AI路标识别算法将“禁止通行”的路标识别成为“限速45”。它与数字世界对抗样本的区别是，物理世界的扰动需要抵抗缩放，裁剪，旋转，噪点等图像变换。

传递性与黑盒攻击：生成对抗样本需要知道AI模型参数，但是在某些场景下攻击者无法得到模型参数。Papernot等人[4]发现对一个模型生成的对抗样本也能欺骗另一个模型，只要两个模型的训练数据是一样的。这种传递性（Transferability）可以用来发起黑盒攻击，即攻击者不知道AI模型参数。其攻击方法是，攻击者先对要攻击的模型进行多次查询，然后用查询结果来训练一个“替代模型”，最后攻击者用替代模型来产生对抗样本。产生出来的对抗样本可以成功欺骗原模型。

3.2 药饵攻击

AI系统通常用运行期间收集的新数据进行重训练，以适应数据分布的变化。 例如，入侵检测系统（IDS）持续在网络上收集样本，并重新训练来检测新的攻击。在这种情况下，攻击者可能通过注入精心设计的样本，即药饵，来使训练数据中毒（被污染），最终危及整个AI系统的正常功能，例如逃逸AI的安全分类等。深度学习的特点是需要大量训练样本，所以样本质量很难完全保证。

Jagielski等人[5]发现，可以在训练样本中掺杂少量的恶意样本，就能很大程度干扰AI模型准确率。他们提出最优坡度攻击、全局最优攻击、统计优化攻击三种药饵攻击。并展示了这些药饵攻击对于健康数据库，借贷数据库跟房价数据库的攻击，影响这些AI模型对新样本的判断。通过加入药饵数据影响对用药量的分析、对贷款量/利息的分析判断、对房子售价的判断。通过加入8%的恶意数据，攻击者能够使模型对超过50%的患者的用药量建议时，出现超过75%的变化量。

3.3 后门攻击

与传统程序相同，AI模型也可以被嵌入后门。只有制造后门的人知道如何触发，其他人无法知道后门的存在，也无法触发。与传统程序不同的是，神经网络模型仅由一组参数构成，没有源代码可以被人读懂，所以后门的隐蔽性更高。攻击者通过在神经网络模型中植入特定的神经元生成带有后门的模型，使得模型虽然对正常输入与原模型判断一致，但对特殊输入的判断会受攻击者控制。如Gu等人[6]提出一种在AI模型中嵌入后门的方法，只有输入图像中包含特定图案才能触发后门，而其他人很难通过分析模型知道这个图案或这个后面的存在。此类攻击多发生在模型的生成或传输过程。

3.4 模型窃取攻击

模型/训练数据窃取攻击是指攻击者通过查询，分析系统的输入输出和其他外部信息，推测系统模型的参数及训练数据信息。与Software-as-a-Service类似，云服务商提出了AI-as-a-Service（AIaaS）的概念，即由AI服务提供商负责模型训练和识别等服务。这些服务对外开放，用户可以用其开放的接口进行图像，语音识别等操作。Tramèr等学者[7]提出一种攻击，通过多次调用AIaaS的识别接口，从而把AI模型“窃取”出来。这会带来两个问题：一是知识产权的窃取。样本收集和模型训练需要耗费很大资源，训练出来的模型是重要的知识产权。二是前文提到的黑盒闪避攻击。攻击者可以通过窃取的模型构造对抗样本。

04 AI安全防御手段

图1描绘了AI系统部署到业务场景中所需要三个层次的防御手段：1、攻防安全：对已知攻击所设计的有针对性的防御机制；2、模型安全：通过模型验证等手段提升模型健壮性；3、架构安全：在AI部署的业务中设计不同的安全机制保证架构安全。

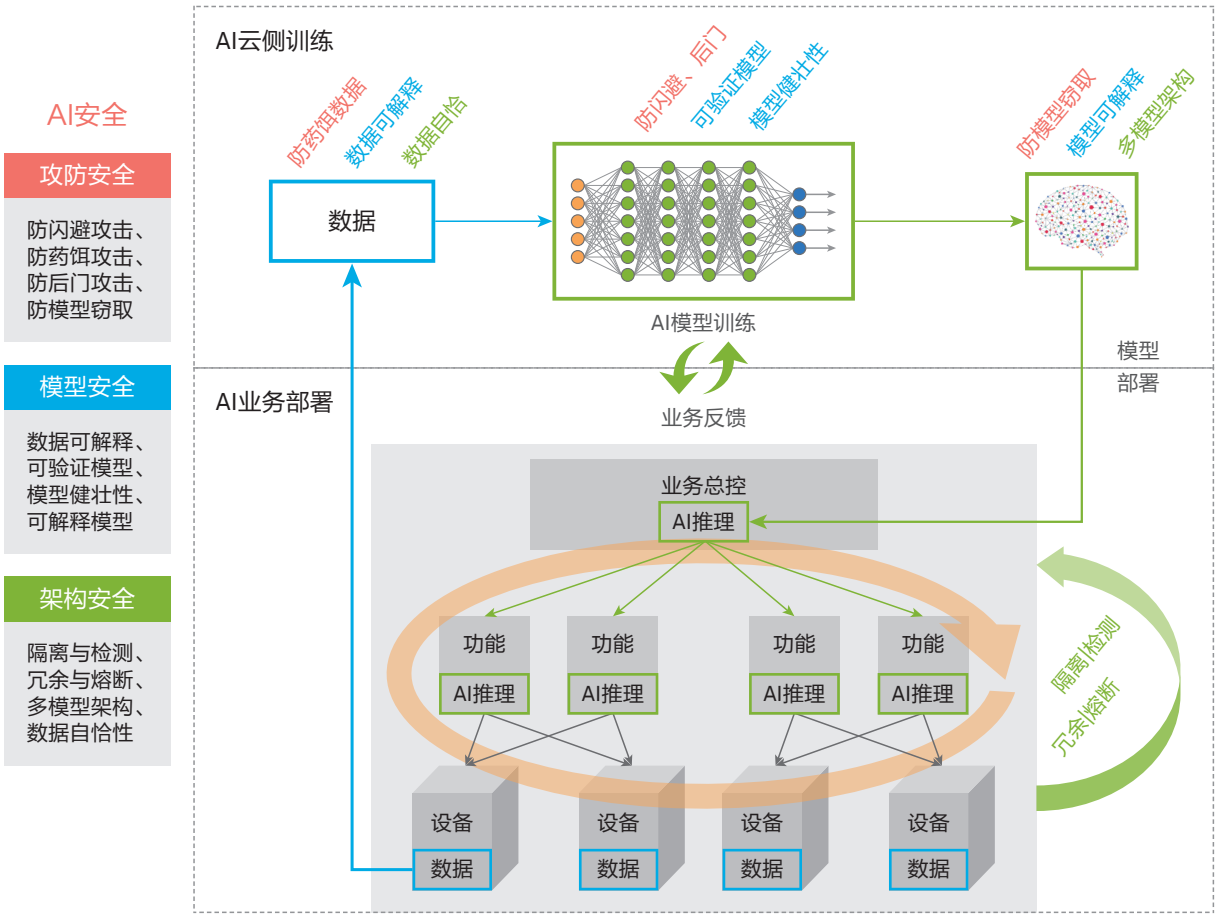


图1 AI安全防御架构

4.1 AI安全攻防

针对上一章提到已知的攻击方式，学术界已有许多对抗方法，对于可能遭受的攻击能提供不同程度的缓解，图2列出AI系统在数据收集、模型训练及模型使用阶段的各种防御技术。

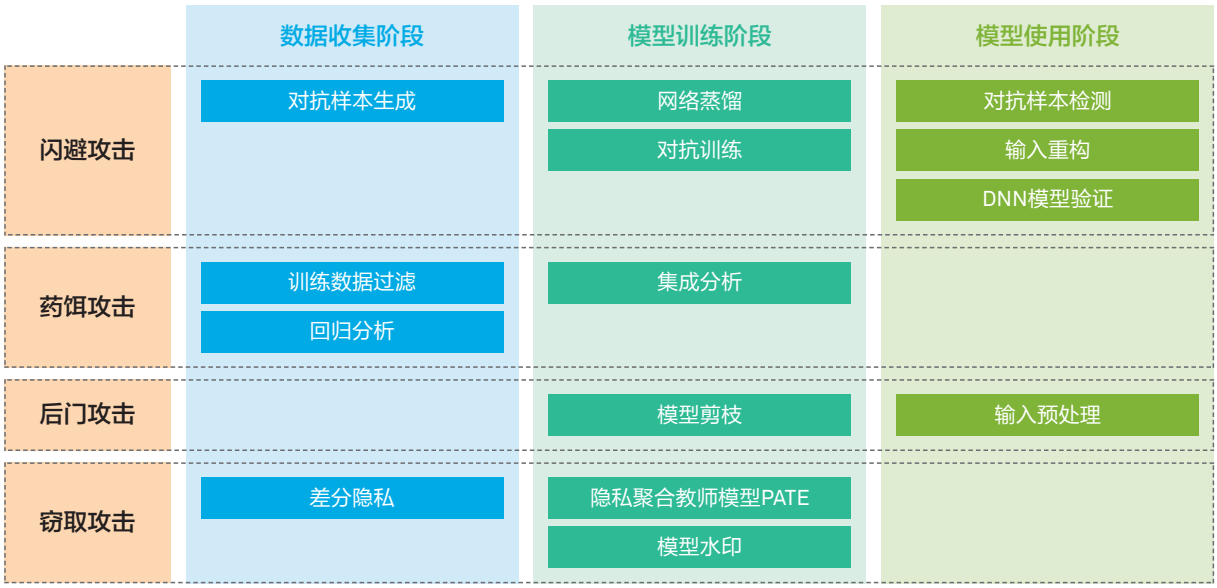


图2 AI安全防御技术

闪避攻击防御技术：

1. **网络蒸馏（Network Distillation）：**网络蒸馏技术的基本原理是在模型训练阶段，对多个DNN进行串联，其中前一个DNN生成的分类结果被用于训练后一个DNN。有学者[8]发现转移知识可以一定程度上降低模型对微小扰动的敏感度，提高AI模型的鲁棒性，于是提出将网络蒸馏技术用于防御闪避攻击，并在MNIST和CIFAR-10数据集上测试，发现该技术可将使特定攻击（如JSMA）的成功率降低。

2. **对抗训练（Adversarial Training）：**该技术的基本原理是在模型训练阶段，使用已知的各种攻击方法生成对抗样本，再将对抗样本加入模型的训练集中，对模型进行单次或多次重训练，生成可以抵抗攻击扰动的新模型。同时，由于综合多个类型的对抗样本使得训练集数据的增多，该技术不但可以增强新生成模型的鲁棒性，还可以增强模型的准确率和规范性。

3. **对抗样本检测（Adversarial Sample Detection）：**该技术的原理为在模型的使用阶段，通过增加外部检测模型或原模型的检测组件来检测待判断样本是否为对抗样本。在输入样本到达原模型前，检测模型会判断其是否为对抗样本。检测模型也可以在原模型每一层提取相关信息，综合各种信息来进行检测。各类检测模型可能依据不同标准来判断输入是否为对抗样本。例如，输入样本和正常数据间确定性的差异可以用来当作检测标准；对抗样本的分布特征，输入样本的历史都可以成为判别对抗样本的依据。

4. **输入重构（Input Reconstruction）：**该技术的原理是在模型的使用阶段，通过将输入样本进行变形转化来对抗闪避攻击，变形转化后的输入不会影响模型的正常分类功能。重构方法包括对输入样本加噪、去噪、和使用自动编码器（autoencoder）[9]改变输入样本等方法。

5. **DNN模型验证（DNN Verification）：**类似软件验证分析技术，DNN模型验证技术使用求解器（solver）来验证DNN模型的各种属性，如验证在特定扰动范围内没有对抗样本。但是通常验证DNN模型是NP完全问题，求解器的效率较低。通过取舍和优化，如对模型节点验证的优先度选择、分享验证信息、按区域验证等，可以进一步提高DNN模型验证运行效率。

以上各个防御技术都有具体的应用场景，并不能完全防御所有的对抗样本。除此之外，也可以通过增强模型的稳定性来防御闪避攻击，使模型在功能保持一致的情况下，提升AI模型抗输入扰动的能力。同时也可以将上述防御技术进行并行或者串行的整合，更有效的对抗闪避攻击。

药饵攻击防御技术：

1. **训练数据过滤（Training Data Filtering）：**该技术侧重对训练数据集的控制，利用检测和净化的方法防止药饵攻击影响模型。具体方向包括[10]：根据数据的标签特性找到可能的药饵攻击数据点，在重训练时过滤这些攻击点；采用模型对比过滤方法，减少可以被药饵攻击利用的采样数据，并过滤数据对抗药饵攻击。

2. **回归分析（Regression Analysis）：**该技术基于统计学方法，检测数据集中的噪声和异常值。具体方法包括对模型定义不同的损失函数（loss function）来检查异常值，以及使用数据的分布特性来进行检测等。

3. **集成分析（Ensemble Analysis）：**该技术强调采用多个子模型的综合结果提升机器学习系统抗药饵攻击的能力。多个独立模型共同构成AI系统，由于多个模型采用不同的训练数据集，整个系统被药饵攻击影响的可能性进一步降低。

此外，通过控制训练数据的采集、过滤数据、定期对模型进行重训练更新等一系列方法，提高AI系统抗药饵攻击的综合能力。

后门攻击防御技术：

1. **输入预处理（Input Preprocessing）：**该方法的目的是过滤能触发后门的输入，降低输入触发后门、改变模型判断的风险[11]。

2. **模型剪枝（Model Pruning）：**该技术原理为适当剪除原模型的神经元，在保证正常功能一致的情况下，减少后门神经元起作用的可能性。利用细粒度的剪枝方法[12]，可以去除组成后门的神经元，防御后门攻击。

模型/数据防窃取技术：

1. **隐私聚合教师模型（PATE）：**该技术的基本原理是在模型训练阶段，将训练数据分成多个集合，每个集合用于训练一个独立DNN模型，再使用这些独立DNN模型进行投票的方法共同训练出一个学生模型[13]。这种技术保证了

学生模型的判断不会泄露某一个特定训练数据的信息，从而确保了训练数据的隐私性。

2. **差分隐私（Differential Privacy）：**该技术是在模型训练阶段，用符合差分隐私的方法对数据或模型训练步骤进行加噪。例如有学者提出使用差分隐私生成梯度的方法[14]，保护模型数据的隐私。

3. **模型水印（Model Watermarking）：**该技术是在模型训练阶段，在原模型中嵌入特殊的识别神经元。如果发现有相似模型，可以用特殊的输入样本识别出相似模型是否通过窃取原模型所得。

4.2 AI模型安全

如上节所述， 恶意机器学习（Adversarial ML）广泛存在，闪避攻击（Evasion）、药饵攻击（Poisoning）以及各种后门漏洞攻击无往不利，攻击不但精准、也有很强的可传递性（Transferability），使得AI模型在实用中造成误判的危害极大。因此，除了针对那些已知攻击手段所做的防御之外，也应增强AI模型本身的安全性，避免其它可能的攻击方式造成的危害，可以由如下图3中列出的几个方面展开。

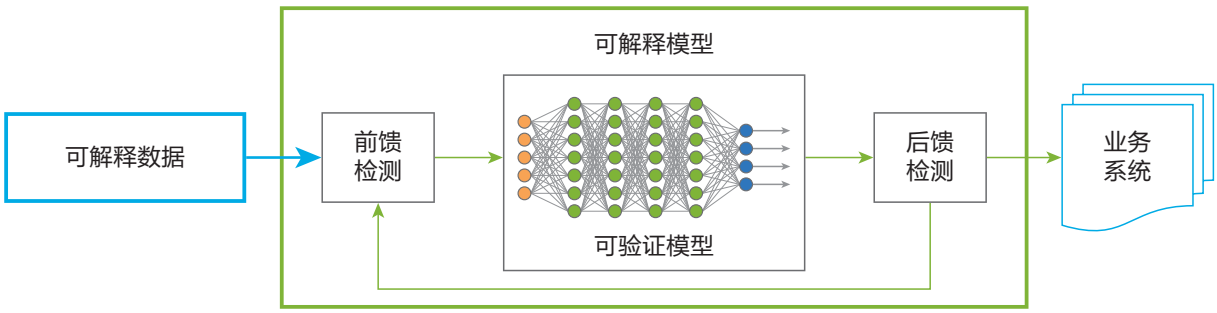


图3 模型安全性分析

模型可检测性：如同传统程序的代码检测，AI模型也可以通过各种黑盒、白盒测试等对抗检测技术来保证一定程度的安全性，已有测试工具基本都是基于公开数据集，样本少且无法涵盖很多其他真实场景，而对抗训练技术则在重训练的过程中带来较大的性能损耗。在AI系统的落地实践中，需要对各种DNN模型进行大量的安全测试，如数据输入训练模型前要做前馈检测模块过滤恶意样本，或模型输出评测结果经过后馈检测模块从而减少误判，才能在将AI系统部署到实际应用前提升AI系统的鲁棒性。

模型可验证性：DNN模型有着比传统机器学习更加预想不到的效果（如更高识别率，更低误报率等），目前广泛用于各种图像识别、语音识别等应用中，然而AI模型在关键安全应用（如自动驾驶、医学诊断等）领域还需要慎重。对DNN模型进行安全验证（certified verification）也可以在一定程度上保证安全性。模型验证一般需要约束输入空间（input space）与输出空间（output space）的对应关系，从而验证输出在一定的范围内。但是基于统计优化（optimization）的学习及验证方法总还是无法穷尽所有数据分布，而极端攻击则有机可乘，这样在实际应用中较难实施具体的保护措施。只有在对DNN模型内部工作机理充分理解的基础上才能进一步解决机制性防御（principled defense）问题。

模型可解释性：目前大多数AI都被认为是一个非常复杂的黑盒子系统，他的决策过程，判断逻辑，判断依据都很难被人完全理解。目前有些业务中，例如棋类、翻译业务，为了让人类和机器之间有更好的互动，我们希望理解为什么

机器做出了这些决定，但是AI系统不可解释并不会带来太多问题。如果它不告诉我们为什么把这个单词翻译成了另一个单词，只要翻译出的结果是好的，它就可以继续是一个完全的黑盒子、完全复杂的系统，而不会带来什么问题。但对于有些业务，不可解释性往往对于会带来业务法务风险或者业务逻辑风险。例如在保险、贷款分析系统中，如果AI系统不能给出其分析结果的依据，那么就有可能会被诟病其带有歧视；又例如在医疗保健中，为了精确的根据AI的分析进行进一步的处理，我们需要了解AI做出判断的根据。例如我们希望AI系统就其判断一位病人有没有癌症给出其数据分析及原因，AI系统需要有能力说“我把这些数据、图像和这个和那个做了对比从而得出了结论”。如果连其运作的原理都无法得知，自然也就无法有效地设计一个安全的模型。增强AI系统的可解释性，都有助于我们分析AI系统的逻辑漏洞或者数据死角，从而提升AI系统安全性，打造安全AI。

学术界正在对AI模型的可解释性进行积极探索，如Strobelt等人[15]提出对隐藏激活函数做可视化分析；Morcos等人[16]提出用统计分析方法发现语义神经元；以及Selvaraju等人[17]提出的针对图形识别的显著性检测。模型可解释性也可以通过以下三个阶段展开：

1. **建模前的“数据可解释”：**模型是由数据训练而来，因此要解释模型的行为，可以从分析训练此模型的数据开始。如果能从训练数据中找出几个具代表性的特征，可以在训练时选择需要的特征来构建模型，有了这些有意义的特征，便可对模型的输入输出结果有较好的解释。

2. **构建“可解释模型”：**一个方法是结合传统机器学习，对AI结构进行补充。这种做法可以平衡学习结果的有效性与学习模型的可解释性，为解决可解释性的学习问题提供了一种框架。传统机器学习方法共同的重要理论基础之一是统计学，在自然语言处理、语音识别、图像识别、信息检索和生物信息等许多计算机领域已经获得了广泛应用并给出很好的可解释性。

3. **对已构筑模型进行解释性分析：**通过分析AI模型的输入、输出、中间信息的依赖关系分析及验证模型的逻辑。学术界中既有如LIME（Local Interpretable Model-Agnostic Explanations）[18]等能够通用地分析多种模型的分析方法，也有需要针对模型构造进行深入分析的分析方法。

当AI系统具有可解释性时，我们就可以比较有效地对系统进行验证和检测：例如通过针对AI系统各模块及输入数据间逻辑关系分析，可以确认客户偿还能力分析模块与客户性别，种族无关。而AI系统具备可解释性的另一个优势是，AI系统的输入/中间数据之间的逻辑关系会相对清晰。我们可以根据这些数据之间的自洽性判断是否有非法/攻击数据，甚至对恶意的攻击样本进行清除跟修复，提高模型健壮性。

欧盟一般数据保护法GDPR要求AI系统决策不能基于如用户种族、政治立场、宗教信仰等数据。而具备可解释性的AI系统可以确保其分析结论符合上述要求，避免出现受到“算法歧视”的受害人。大多AI系统中，其偏见问题往往不在于算法本身，而是提供给机器的数据。如果输入数据中带有存在偏见的数据，例如公司HR有轻微拒绝女性求职者的偏见，这些数据将导致模型中的拒绝女性求职者案例增加，从而造成性别比例失调。即使性别并不是模型培训数据的重要特征，其数据也会使AI模型的分析结论进一步放大人类的本身偏见。而政府往往需要验证AI使能系统的安全性，可靠性，可解释性。只有可解释，可验证的健壮AI系统才能给予公众信心与信任。

4.3 AI业务的安全架构

在大力发展人工智能的同时，必须高度重视AI系统引入可能带来的安全风险，加强前瞻预防与约束引导，最大限度降低风险，确保人工智能安全、可靠、可控发展。而在业务中使用AI模型，则需要结合具体业务自身特点和架构，分析判断AI模型使用风险，综合利用**隔离、检测、熔断和冗余**等安全机制设计AI安全架构与部署方案，增强业务产品健壮性。

在自动驾驶业务中，当AI系统如果对刹车，转弯，加速等等关键操作的判断出现失误时，可能会对用户，对社会造成巨大危害。因此需要保证AI系统在关键操作时的安全使用。对自动驾驶AI系统进行许多的安全测试当然很重要，但是这种模拟测试方法并不能保证AI系统不出错。在很多业务中，也许很难找到一个任何时候都能给出100%正确答案的AI系统。相比之下，更重要的是对系统架构进行安全设计，使得当AI系统对判断不确定的时候，业务还能够回退到手工操作等安全状态。在医疗辅助AI系统中，如果AI系统对于“应该给病人哪个药，用量多少”这个问题不能给出确定答案时，或感知到自身有可能受到攻击时，相比给出一个可能造成危险的不准确预测，让AI系统直接回答“请咨询病人的医师”会更好一点。为了保护用户利益，我们需要按照业务需求，在系统中合理运用如下安全机制确保AI业务安全，如图4所示：

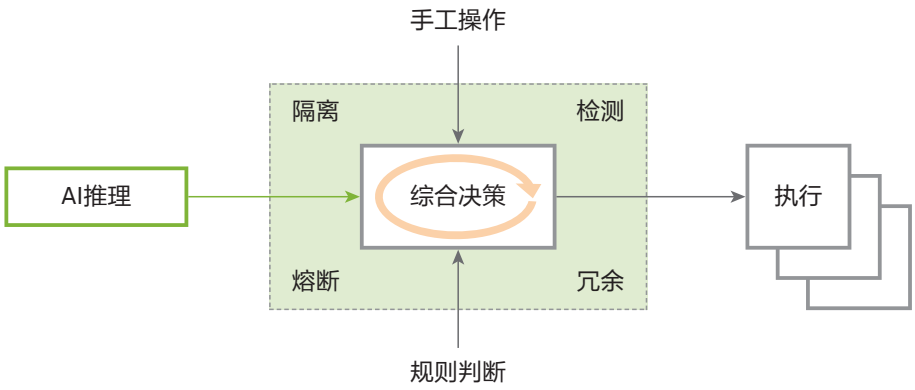


图4 AI引入业务决策的安全架构

1. **隔离：**在满足业务稳定运行的条件约束下，AI系统会分析识别最佳方案然后发送至控制系统进行验证并实施。通常业务安全架构要考虑对各个功能模块进行隔离，并对模块之间设置访问控制机制。对AI系统的隔离可以一定程度上减少针对AI推理的攻击面，而对综合决策系统的隔离可以有效减少针对决策系统的攻击。AI推理的输出作为辅助决策建议将导入综合决策模块，而只有经过授权认证的指令才能得以通过。

2. **检测：**在主业务系统中部署持续监控和攻击检测模型，综合分析网络系统安全状态，给出系统当前威胁风险级别。当威胁风险较大时，综合决策可以不采纳自动系统的建议，而是将最终控制权交回人员控制，保证在遭受攻击情况下的安全性。

3. **熔断：**业务系统在进行关键操作时，如AI辅助的自动驾驶或医疗手术等，通常要设置多级安全架构确保整体系统安全性。需要对AI系统给出的分析结果进行确定性分析，并在确定性低于阈值时回落到以规则判断为准的常规技术或直接交回人工处理。

4. **冗余：**很多业务决策、数据之间具有关联性，一个可行的方法是通过分析此类关联性是否遭受破坏保证AI模型运行时的安全。还可以搭建业务“多模型架构”：通过对关键业务部署多个AI模型，使得在单个模型出现错误时不会影响到业务最终决策。同时多个模型的部署也使得系统在遭受单一攻击时被全面攻克的可能性大大降低，从而提升整个系统的强壮性。

Amodei等人[19]还进一步描述了AI系统在应用中可能会遇到的几种安全挑战：如避免AI系统在执行任务时可能产生的消极副作用、AI系统在达成目的时可能采取的趋利行为、以及AI系统在执行任务时的安全拓展问题等。对这些问题进行基础研究将会使得AI系统在未来实用场景更加安全。

05 携手共建安全的智慧未来

人工智能的各个学科，如计算机视觉、语音识别、自然语言处理、认知与推理、博弈等，还处在早期发展的阶段，依靠大数据做统计分析的深度学习系统拓展了人工智能所能解决问题的边界，但也被认为是普遍“缺乏常识”，这也是当前人工智能研究的最大障碍。人工智能要依靠数据与知识的双轮驱动，下一代人工智能的突破可能是知识推理。而人工智能应用的大规模普及和发展则需要很强的安全性保证。我们首先关注两大类AI安全攻防问题：第一类是攻击者影响AI决策的正确性：攻击者可以通过破坏和控制AI系统本身，或者通过特意改变输入来使系统不知不觉地做出攻击者想要的决定；第二类是攻击者获取AI系统训练的保密数据，或者破解AI模型。本文进一步从AI安全攻防、AI模型安全和AI架构安全等三个层面阐述AI系统安全，保障AI应用的安全性。此外，AI的透明性和可解释性也是安全的基础，一个不透明和无法解释的人工智能无法承担起涉及人身安全及公共安全的关键任务。

人工智能还会带来法律法规、伦理道德、社会监管等很宽泛的安全课题。2016年9月1日，斯坦福大学“**人工智能百年研究**（AI100）”项目发布了首篇名为“2030 年的人工智能与生活（AI and Life in 2030）”研究报告[20]，指出面对人工智能技术将带来的深刻变化，要求更合理和“不会扼杀创新”的监管。未来几年，随着人工智能在交通和医疗等领域内的应用，它们必须以一种能构建信任和理解的方式引入，还要尊重人权和公民权利。与此同时，“政策和流程也应该解决道德、隐私和安全方面的影响”。为此国际社会应协同合作推动人工智能向着造福人类的方向演进。

参考文献

[1] I. Stoica, D. Song, R. A. Popa, D. Patterson, M. W. Mahoney, R. Katz, A. D. Joseph, M. Jordan, J. M. Hellerstein, J. Gonzalez, K. Goldberg, A. Ghodsi, D. Culler and P. Abbeel, "A Berkeley View of Systems Challenges for AI," University of California, Berkeley, Technical Report No. UCB/EECS-2017-159, 2017.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[3] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno and D. Song, "Robust physical-world attacks on deep learning models," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] N. Papernot, P. McDaniel and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," arXiv preprint arXiv:1605.07277, 2016.

[5] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *IEEE Symposium on Security and Privacy (S&P)*, 2018.

[6] T. Gu, B. Dolan-Gavitt and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," in *NIPS MLSec Workshop*, 2017.

[7] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter and T. Ristenpart, "Stealing Machine Learning Models via Prediction APIs," in *USENIX Security Symposium*, 2016.

[8] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks,," in *IEEE Symposium on Security and Privacy (S&P)*, 2016.

[9] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[10] R. Laishram and V. Phoha, "Curie: A method for protecting SVM classifier from poisoning attack," arXiv preprint arXiv:1606.01584, 2016.

[11] Y. Liu, X. Yang and S. Ankur, "Neural trojans," in *International Conference on Computer Design (ICCD)*, 2017.

[12] K. Liu, D.-G. Brendan and G. Siddharth, "Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks," arXiv preprint arXiv:1805.12185, 2018.

[13] N. Papernot, A. Martín, E. Ulfar, G. Ian and T. Kunal, "Semi-supervised knowledge transfer for deep learning from private training data," arXiv preprint arXiv:1610.05755, 2016.

[14] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar and L. Zhang, Deep learning with differential privacy, ACM SIGSAC Conference on Computer and Communications Security, 2016.

[15] H. Strobelt, S. Gehrmann, H. Pfister and A. M. R. Lstmvis, "A tool for visual analysis of hidden state dynamics in recurrent neural networks," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 667-676, 2018.

[16] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz and M. Botvinick, "On the importance of single directions for generalization," arXive preprint arXiv:1803.06959, 2018.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," arXiv preprint arXiv:1610.02391, 2016.

[18] M. T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *ACM international conference on knowledge discovery and data mining (KDD)*, 2016.

[19] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, "Concrete Problems in AI Safety," arXiv preprint arXiv:1606.06565, 2016.

[20] S. Peter, B. Rodney, B. Erik, C. Ryan, E. Oren, H. Greg, H. Julia, K. Shivaram, K. Ece, K. Sarit, L.-B. Kevin, P. David, P. William, S. AnnaLee, S. Julie and etc, Artificial Intelligence and Life in 2030, One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, 2016.

商标声明


、**HUAWEI**、**华为**、 
是华为技术有限公司的商标或者注册商标。

在本手册中以及本手册描述的产品中，出现的其他商标、产品名称、服务名称以及公司名称，由其各自的所有人拥有。

免责声明

本文档可能含有预测信息，包括但不限于有关未来的财务、运营、产品系列、新技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何要约或承诺。华为可能不经通知修改上述信息，恕不另行通知。

华为技术有限公司

深圳市龙岗区坂田华为基地

电话：(0755) 28780808

邮编：518129

www.huawei.com